Improving sparse recovery on structured images with bagged clustering

Andrés HOYOS-IDROBO*, Yannick SCHWARTZ*, Gaël VAROQUAUX*, Bertrand THIRION*,
*INRIA Parietal, Neurospin, bât 145, CEA Saclay, 91191 Gif sur Yvette, France
firstname.lastname@inria.fr

***Abstract*—The identification of image regions associated with external variables through discriminative approaches yields ill-posed estimation problems. This estimation challenge can be tackled by imposing sparse solutions. However, the sensitivity of sparse estimators to correlated variables leads to non-reproducible results, and only a subset of the important variables are selected. In this paper, we explore an approach based on bagging clustering-based data compression in order to alleviate the instability of sparse models. Specifically, we design a new framework in which the estimator is built by averaging multiple models estimated after feature clustering, to improve the conditioning of the model. We show that this combination of model averaging with spatially consistent compression can have the virtuous effect of increasing the stability of the weight maps, allowing a better interpretation of the results. Finally, we demonstrate the benefit of our approach on several predictive modeling problems.**

***Keywords*-Stability; High-dimensional estimators; machine learning; brain imaging; clustering**

I. Introduction

Using machine learning on neuroimaging data, brain regions can be linked with external variables[1]. In particular, linear predictive models are interesting as their coefficients form brain maps that can be interpreted. However, because of the high dimensionality of brain imaging data, their estimation is an ill-posed problem and in order to find a feasible solution, some constrains must be imposed to the estimator. A popular way to solve that problem is to use a sparsity constraint as it isolates putatively relevant features. In practice, the high correlation between neighboring voxels leads to selecting too few features, and hinders the estimators' ability to recover a stable support. The estimation instability causes a high variance of both the prediction scores and the model coefficients, and therefore may result in non reproducible findings [2], [3]. To mitigate this instability, *stability selection* [2] adds randomization to sparsity for feature selection. More generally, two main classes of methods are known to stabilize models with high-correlated features. *Feature clustering based methods* reduce both the local correlations and the dimensionality of the data [4], [5]. *Model aggregation methods* –such as bagging– generate new training sets from an original one with data perturbation schemes. They build multiple estimators from the perturbed data to combine them in an estimate with reduced variance. Stability selection methods [2], [6] are variants that use sparsity to perform feature selection. In [5], the authors found that, for fMRI data, combining clustering with stability-selection method can select a small number of voxels that increases the prediction score and enhances the interpretability of the weight maps.

Our contribution: Clustering reduces the local correlation of the voxels as well as the dimensionality of the data at the expense of spatial resolution. We propose to combine it with bagging to mitigate the loss of spatial resolution. Importantly, we do not rely on feature selection but on model averaging, which is an easier problem and enables us to cut computation times. We show that our method outperforms standard sparse classifiers in prediction accuracy, feature recovery, and computational efficiency.

Our method relates to the model proposed in [5] but differs in the following: i) we rely here on bagging rather than voxel selection, and ii) our model estimates far fewer classifiers, which makes it computationally more efficient.

II. Methods

Sparse recovery: Consider the linear regression model

$$y = f\left(\mathbf{X}\beta + \epsilon\right), \tag{1}$$

where $f(.)$ represents the decision function, $y \in \mathbb{R}^n$ is the categorical/behavioral variable related to the experimental condition, $\mathbf{X} \in \mathbb{R}^{n \times p}$ represents n observed brain images composed of p voxels, and ϵ is a noise vector where $\epsilon \sim \mathcal{N}(0, \sigma^2 I_n)$. Given a pair $(y, \mathbf{X})$ the goal is to estimate the unknown weight vector $\beta \in \mathbb{R}^p$. Typically, n is a few hundreds of images while p can be thousands of voxels ($p \gg n$). Due its high-dimensional and ill-conditioned nature, the model (1) is not identifiable. This challenge can be tackled by imposing a sparse estimate for β, e.g. by seeking a small number of non-zero β-coefficients. The standard approach relies on using ℓ_1 penalization:

$$\hat{\beta}(\lambda) = \underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \left(\mathcal{L}(y, \mathbf{X}, \beta) + \lambda \|\beta\|_1\right), \tag{2}$$

where $\lambda \in \mathbb{R}^+$ is the regularization parameter that controls the amount of sparsity in the estimate $\hat{\beta}$, and $\mathcal{L}(y, \mathbf{X}, \beta)$ is the loss function that measures the quality of the estimator. Here, we focus on the logistic loss, $\log(1 + \exp(-y\mathbf{X}\beta))$, given that it is well-suited for classification settings. The corresponding classifier is a sparse logistic regression (SLR), used routinely in neuroimaging [7].

Formal conditions for success in the recovery of an s-sparse vector require the irrelevant variables not to be too correlated with the relevant ones. This is known as Incoherence or *Irrepresentability* criterion [8]. In practice, if several columns of $\mathbf{X}$ are strongly correlated, sparse estimators will select arbitrarily only one of them, leading for instance to a high rate of false negatives in the support estimation (see Fig. 4). Below, we address this issue, using clustering techniques.

Clustering: Clustering techniques have been frequently used in medical imaging as a means to compress information, with empirical success [9]. Indeed, they account for the latent structure of these images. The aim of clustering methods is to define a suitable partition of the image volume, where local averages of the signal are grouped, capturing the local correlations into spatial clusters. By doing this, the high-frequency noise is reduced while preserving the low-frequency signal of interests. Let $(u_i)_{i\in[q]}$ be the set of projectors to q spatial components (clusters), the reduced design $\mathbf{X}_{\text{red}} \in \mathbb{R}^{n\times q}$ is defined as

$$\mathbf{X}_{\text{red}} = RNN(\mathbf{X}, q) = \langle \mathbf{X}, u_i \rangle \,,\ \forall i \in [q]. \tag{3}$$

The ensuing estimates can then be embedded back into the original space (lossy compression) by *inverting* the projectors as a piecewise constant mapping, yielding $\mathbf{X}_{comp} \in \mathbb{R}^{n\times p}$. $\mathbf{X}_{\text{red}}$ is expected to yield better-behaved estimates, as there are less variables and less correlation between them. Data clustering can be performed through various approaches, such as k-means or agglomerative clustering. Here we use a recursive nearest neighbor agglomeration (RNN) taking into account the 3D image lattice structure.

Bagging multiple clusterings: As a given clustering is a lossy representation of the image domain, we use multiple clusterings and average the results across them. For this purpose, we sub-sample the data, resorting to standard K-fold approaches as in cross-validation: each clustering is slightly different in each fold[1]. On each fold, we estimate a linear model with the corresponding clustering. We finally average models across folds. Thus the different clustering solutions complete each other and each weight map does not reflect only one given set of clusters. The benefit of using a K-fold approach is also that we can use the out-of-bag prediction error to set the hyper-parameters of the estimator: in each fold, we compute a regularization path, and select on this path the estimator that minimizes the error on left-out data. A summary of this method is presented in the algorithm 1.

Algorithm 1 Sparse Clustered Logistic Regression (SCLR)

Require: Input data $\mathbf{X}$ with shape (p, n), q number of clusters, $[\lambda_1, \ldots, \lambda_M] \in \Lambda \subset \mathbb{R}^+$ the range of values for the sparsity-inducing penalty, J the number of resampling sets k_j, sub-sampling fraction $\pi \in [0, 1]$.

for $j \leftarrow 1$ **to** J **do**
  $\tilde{\mathbf{X}} \leftarrow \mathbf{X}[k_j] \in \mathbb{R}^{\pi n\times p}$ (Sub-sample)
  $\mathbf{X}_{\text{red}} \leftarrow RNN\left(\tilde{\mathbf{X}}[k], q\right)$, $\mathbf{X}_{\text{red}} \in \mathbb{R}^{\pi n\times q}$ (Image reduction)
  for $m \leftarrow 1$ **to** M **do**
    $\hat{\beta}^m_{\text{red}}(\lambda_m) \leftarrow \underset{\beta\in\mathbb{R}^q}{\arg\min} \{\mathcal{L}(y, \mathbf{X}_{\text{red}}; \beta) + \lambda_m \|\beta\|_1\}$ (Solve the ℓ_1 logistic regression)
  end for
  $\hat{\beta}^j_{\text{best-red}} \leftarrow \mu\left((\hat{\beta}^i_{\text{red}})_{i\in[1,\ldots,J]}\right)$ (μ selects the best estimator)
  $\hat{\beta}^j_{\text{best}} \leftarrow RNN^{-1}(\hat{\beta}^k_{\text{best-red}})$ (Returning to the voxels space)
end for
$\hat{\beta}_{\text{agg}} \leftarrow J^{-1}\sum_{j=1}^{J} \hat{\beta}^j_{\text{best}}$ (Mean aggregation of the best models)
return Solution $\hat{\beta}_{\text{agg}}$

III. Experiments

As our aim is to assess model stability and prediction accuracy, we directly turn to a series of discriminative tasks involving public neuroimaging datasets (either anatomical or functional) to evaluate how reproducible the predictions and coefficients obtained through SLR/SCLR are.

functional Magnetic Resonance Imaging (fMRI).: We use 4 studies drawn from the OpenfMRI[2] project [10]. Specifically, we relied on the object recognition tasks and cueing tasks. Contrasts of these tasks were obtained by general linear model application upon the preprocessed data resampled at 3mm resolution in the MNI space, which yields about $54,239$ voxels. We focus here on the binary classification problem that consists in predicting which task the subject was performing (e.g. viewing a cat or a chair) in an inter-subject discrimination setting.

Anatomical: We performed a discriminative task on the OASIS dataset [11]: prediction of the gender of the subject. We used 403 anatomical images and processed them with the SPM8 software to obtain VBM modulated grey matter density maps sampled in the MNI space at 2mm resolution. The images were masked to an approximate average mask of the grey matter, leaving $140,398$ voxels.

For the training step, we leave half of the subjects out, so that the other half remains unseen by the estimator. To assess the stability of the model, we create a set of artificial data from the training data using 80 Bootstrap samplings, then the estimator is trained and tested on the unseen test data. When clustering (reduction) is applied, the number of clusters is defined as $q \approx 10\%\ p$, this number of clusters represents a good trade-off between resolution and compression on

[1]In practice, when performing clustering, we have found it useful to further sub-sample the observations, to increase the spatial entropy of clusters across different folds.

[2]https://openfmri.org/data-sets

brain images, and with a sub-sample fraction of $\pi = 0.6$ (this number can be chosen arbitrarily in $]0, 1[$). The inner fold is a Leave-One-Subject-Out cross validation. The set of regularization parameters Λ is defined as a grid of 24 values in the log-scale $[10^{-5}, 10^{5}]$ interval.

Implementation aspects: The data that we used are the publicly available OASIS and OpenfMRI datasets. We relied on the scikit-learn library [12] (v0.15) for machine learning tasks (logistic regression) and for clustering. We relied on the Nilearn library for interaction on neuroimaging data.

IV. Results

We benchmark our novel approach against a standard sparse logistic regression, by considering its impact on accuracy and the stability of the brain maps. In the following, *SLR* denotes the sparse logistic regression and *SCLR* denotes the sparse clustered logistic regression.

Prediction accuracy: A scatterplot of the F_1 scores (i.e. harmonic mean of precision and recall) computed over 17 binary classification tasks is presented in Fig. 1. Average scores are given in table I. This shows that the *SCLR* often obtains a higher F_1 score than the classical *SLR*.

Stability: Fig.3 shows the dice coefficient used to measure the spatial overlap between the most representative coefficients of the linear estimator across different bootstrap realizations. We observe that the bagged clustering approach increases the overlap of the weight maps, going from 10% to 16% in average, and contributes to the stabilization of the weights by improving the conditioning of the model and reduction of statistical fluctuations. Fig.2 shows the Z-scores of the maps across bootstrap replicates for two discriminative tasks. We can observe that bagged clustering increases the stability of the weights through a reduction of their variance. In Fig 4 we can see that the bagged clustering approach effectively mitigates the loss of spatial resolution produced by sparse models, increasing the number of true positive detections and leading to a better feature recovery.

Finally, using cluster-based methods to reduce the dimensionality of the problem yields better classification performance on the reduced space than the whole volume with a much smaller computation time (by more than one order of magnitude). For instance, on the *ds108* dataset, we observe a 15-fold speed increase: *SLR* $= 7.64s$, *SCLR*$= 0.49s$.

V. Discussion

We have devised a strategy for enhancing the prediction scores and stability of sparse linear estimators. We built an ensemble of feasible models that maximizes the explainable variance for each fold of cross-validation and decreases collinearity effects by clustering. This model comes with significantly reduced computational cost with respect to the standard sparse classifiers, as it performs the model fitting on a smaller dimensionality and model averaging during the cross-validation step. While there is no guarantee that the spatial clustering is meaningful per se, it brings a good dimensionality reduction that minimizes information loss and captures better the signal than noise. Our experiments demonstrate the ability of the averaged estimator to achieve higher or equal prediction scores in less computational time. But, more importantly, it enhances the interpretability of the results by reducing the variance of the resulting patterns.

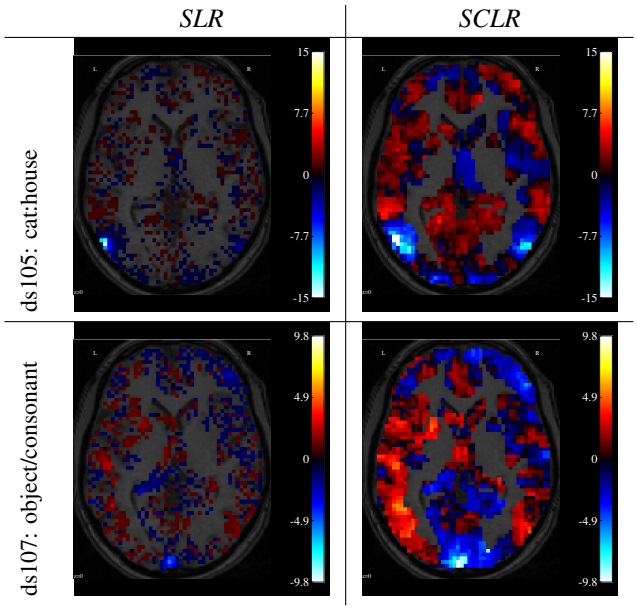

Figure 2. Z-score obtained across bootstraps for two discriminative tasks, using the candidate approaches. Higher values hint at lower variability across bootstrap replications. *SCLR* decreases the variability and yields larger standardized effects.

Acknowledgment: Data were provided in part by the OpenfMRI Project and OASIS project. The OpenfMRI project is managed by Russ Poldrack at the University of Texas at Austin, with computing resources provided by the Texas Advanced Computing Center. It is funded by a grant from the National Science Foundation (OCI-1331441). The OASIS project was supported by grants P50 AG05681, P01 AG03991, R01 AG021910, P50 MH071616, U24 RR021382, R01 MH56584.

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 604102 (Human Brain Project).

References

[1] J. Mourão-Miranda, A. L. Bokde, C. Born, H. Hampel, and M. Stetter, "Classifying brain states and determining the discriminating activation patterns: Support vector machine on functional MRI data," *NeuroImage*, vol. 28, no. 4, pp. 980 – 995, 2005.

[2] N. Meinshausen and P. Bühlmann, "Stability selection," *J Roy Stat Soc B*, vol. 72, no. 4, pp. 417–473, 2010.

[3] B. Yu, "Stability," *Bernoulli*, vol. 19, p. 1484, 2013.

Dataset		*SLR*	*SCLR*
DS105 (Haxby)	bottle/scramble	0.591	**0.626**
	cat/chair	0.558	**0.612**
	cat/house	0.698	**0.963**
	chair/house	0.668	**0.734**
	chair/scramble	0.700	**0.743**
	face/house	**0.766**	0.742
	tools/scramble	0.666	**0.743**
DS107	consonant/scramble	0.886	**0.897**
	objects/consonant	0.855	**0.901**
	objects/scramble	0.863	**0.898**
	objects/words	0.689	**0.708**
	words/scramble	0.782	**0.841**
DS108	negative cue / neutral cue	0.444	**0.497**
	negative rating / neutral rating	0.520	**0.537**
	negative stim / neutral stim	0.734	**0.743**
DS 109	false picture / false belief	0.664	**0.675**
Oasis (VBM)	Gender discrimination	0.617	**0.655**

Table I
AVERAGE PREDICTION SCORES FOR TWO APPROACHES STUDIED, ACROSS 17 CLASSIFICATION PROBLEMS. THE *SCLR* SYSTEMATICALLY INCREASES THE PREDICTION SCORE.

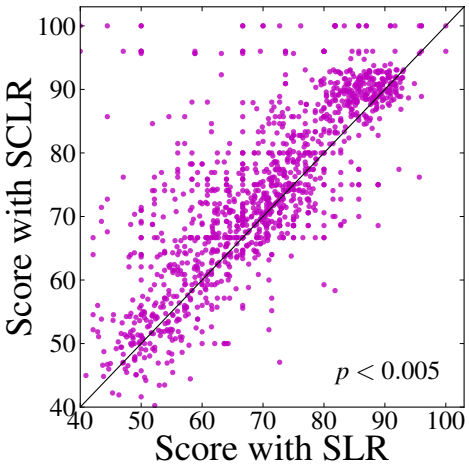

Figure 1. Comparison of the prediction accuracy across approaches: each point represents the F_1-score obtained for a given discrimination task and Boostrap sampling in *SCLR* vs *SLR*. The observed difference is significant ($p < 0.005$, paired Wilcoxon rank test).

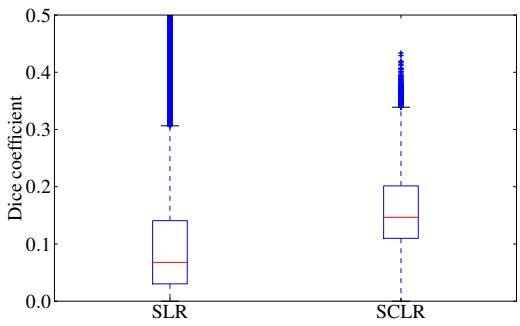

Figure 3. Spatial overlap of the maps, thresholded at 95% and obtained through bootstrap sampling (80 replications), in 17 classification problems. The boxplot represents the $\frac{80\times79}{2}\times 17$ Dice coefficients obtained through the different approaches. This shows that the bagged clustering approach increases the stability of the weight maps in average.

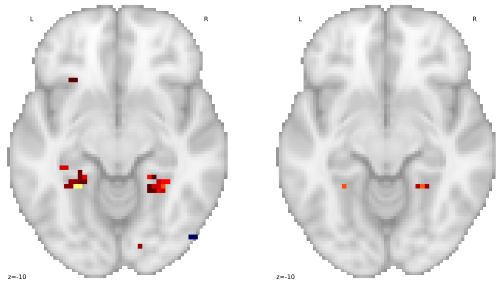

Figure 4. Weight maps obtained for *ds105* dataset and the cat/house classification task, using the two different approaches:(left) averaged clustering (*SCLR*), (right) the standard sparse logistic regression (*SLR*). We can see that bagged clustering increases the feature recovery.

[4] B. D. Mota, V. Fritsch, G. Varoquaux *et al.*, "Enhancing the reproducibility of group analysis with randomized brain parcellations," *MICCAI*, vol. 16, p. 591, 2013.

[5] G. Varoquaux, A. Gramfort, and B. Thirion, "Small-sample brain mapping: sparse recovery on spatially correlated designs with randomization and clustering," in *ICML*, 2012, p. 1375.

[6] J. M. Rondina, J. Shawe-Taylor, and J. Mourão-Miranda, "Stability-based multivariate mapping using scors," in *PRNI*, 2013, p. 198.

[7] S. Ryali, K. Supekar, D. Abrams, and V. Menon, "Sparse logistic regression for whole-brain classification of fMRI data," *NeuroImage*, vol. 51, p. 752, 2010.

[8] N. Meinshausen and B. Yu, "Lasso-type recovery of sparse representations for high-dimensional data," *The Annals of Statistics*, vol. 37, p. 246, 2009.

[9] V. Michel, A. Gramfort, G. Varoquaux *et al.*, "A supervised clustering approach for fMRI-based inference of brain states," *Pattern Recognition*, vol. 45, p. 2041, 2012.

[10] R. A. Poldrack, D. M. Barch, J. P. Mitchell *et al.*, "Toward open sharing of task-based fMRI data: the OpenfMRI project," *Frontiers in Neuroinformatics*, vol. 7, 2013.

[11] D. S. Marcus, T. H. Wang, J. Parker *et al.*, "Open access series of imaging studies (oasis): cross-sectional mri data in young, middle aged, nondemented, and demented older adults." *J Cogn Neurosci*, vol. 19, p. 1498, 2007.

[12] F. Pedregosa, G. Varoquaux, A. Gramfort *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, p. 2825, 2011.